Unison: An Integrated Platform for Computational Biology Discovery

Freely accessible and available at http://unison-db.org/.

Reece Hart, Kiran Mukhyala
Genentech, Inc.

Pacific Symposium on Biocomputing 2009
Sequence Analysis

- i.e., show predictions for a given sequence
- Typically involves minutes to hours of computing per sequence.

Prediction results

- method-specific data such as score, e-value, p-value, kinase probability, etc.

parameters

- execution arguments/options for every prediction type and result

Feature-Based Mining

- i.e., show sequences that contain specified features
- Typically entails days to months of computing results.

sequences

- non-redundant superset of all sequences

feature types/models

- HMM, TM, signal, etc.
Unison in a Nutshell

Protein Sequences and Annotations
Genomes, Gene Mapping & Structure, Probes
Domain, Structure & Homology Predictions
Structures & Ligands
Auxiliary Annotations GO, RIF, SCOP, etc.

Sequences and Annotations
UniProt, IPI, Ensembl, RefSeq, PDB STRING, PHANTOM, HUGE, ROUGE, MGC, Derwent, pataa, nr, etc.
>13M seqs, >17k species, 69 origins

Auxiliary Data
HomoloGene, Gene Ontology, taxonomy, PDB, HUGO, SCOP, etc.

Precomputed predictions
Domains, homology, structure, TMs, localization, signals, disorder, etc.
>200M predictions, 23 types, ~6 CPU-years
Unison has many applications.

- **Unison Web Tools**
- **Other In-House Tools**
- **Ad Hoc Mining**

**Mining and analysis projects**

**Protein Sequences and Annotations**
- Genomes, Gene Mapping & Structure, Probes
- UniProt, IPI, Ensembl, RefSeq, PDB STRING, PHANTOM, HUGE, ROUGE, MGC, Derwent, pataa, nr, etc. >13M seqs, >17k species, 69 origins

**Auxiliary Data**
- HomoloGene, Gene Ontology, taxonomy, PDB, HUGO, SCOP, etc.

**Precomputed predictions**
- Domains, homology, structure, TM, localization, signals, disorder, etc. >200M predictions, 23 types, ~6 CPU-years

**Sequences and Annotations**
- GO, RIF, SCOP, etc.
### Annotation Results

<table>
<thead>
<tr>
<th>Query</th>
<th>Unison pseq id</th>
<th>NCBI Gene &amp; RefSeq</th>
<th>Genes Genes</th>
<th>Cytoband</th>
<th>Probes</th>
<th>GO Functions</th>
<th>Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCL2 HUMAN UniProtKB/Swiss-Prot Apoptosis regul...</td>
<td>Unison:30966</td>
<td>GeneID:596©</td>
<td>RefSeq:NP_000624.2©</td>
<td></td>
<td>18q21.33</td>
<td>See all 25 functions</td>
<td>BH4(7,33,57,5,6e-14), BH4(110-30), BH3(93-107), Bcl-2(197-195,197,5,1e-56), BH1(137-155), BH2(186-199), TM1(214-236)</td>
</tr>
<tr>
<td>BCLX HUMAN UniProtKB/Swiss-Prot Apoptosis regul...</td>
<td>Unison:30974</td>
<td>GeneID:598©</td>
<td>RefSeq:NP_612815.1©</td>
<td>UNQ2707©, PR0107805©, PR011697©, PR0122026©</td>
<td>20q11.21</td>
<td>See all 7 functions</td>
<td>BH4(1-27,59,3e-14), BH4(14-24), BH3(85-100), Bcl-2(90-188,207,5.2e-50), BH1(130-148), BH2(181-192), TM1(213-236)</td>
</tr>
<tr>
<td>MCL1 HUMAN</td>
<td>Unison:2104</td>
<td>GeneID:417©</td>
<td>RefSeq:NP_200796©</td>
<td>UNQ2334©</td>
<td>1q21.2</td>
<td>See all 8 functions</td>
<td>BH3(208-223)</td>
</tr>
</tbody>
</table>
Unison is a platform for diverse tools.

Matt Brauer
Guy Cavet
Josh Kaminker
Scott Lohr
Kathryn Woods
Jean Yuan
Peng Yue
Mining for TNF ligands
Mining for E3 Ligases
Mining for 4H Cytokines
Mining for ITxM
Mining for deubiquitinases
Analyzing SNP impact on binding interfaces
Mining for ITIMs the old way.

➢ Collect sequences.
➢ Prune redundant sequences. (How?!) 
➢ For each unique sequence, predict
  • Immunoglobulin domains.
  • Transmembrane domains.
  • ITIM domains.
➢ Write a program that filters predictions.
➢ Summarize hits with external data.
➢ Do it again when source data are updated.
SELECT IG.pseq_id,
    IG.start as ig_start, IG.stop as ig_stop, IG.score, IG.eval,
    TM.start as tm_start, TM.stop as tm_stop,
    ITIM.start as itim_start, ITIM.stop as itim_stop
FROM pahmm_current_pfam_v IG
JOIN pftmhmm_tms_v TM ON IG.pseq_id = TM.pseq_id AND IG.stop < TM.start
JOIN pfregexp_v ITIM ON TM.pseq_id = ITIM.pseq_id AND TM.stop < ITIM.start
WHERE IG.name = 'ig' AND IG.eval < 1e-2
    AND ITIM.acc = 'MOD_TYR_ITIM';
“Are you sure about this Stan? It seems odd that a pointy head and a long beak is what makes them fly.”

Kiran Mukhyala

Fernando Bazan, Matt Brauer, Jason Hackney, Pete Haverty, Ken Jung, Josh Kaminker, Nandini Krishnamurthy, Li Li, Yun Li, Shiuh-ming Loh, Jinfeng Liu, Peng Yue, Jianjun Zhang, Yan Zhang

http://unison-db.org/
Open access web site, downloads, documentation, references

unison-db.org:5432
PostgreSQL & odbc/jdbc/sdbc access
**Unison Contents**

- **patents**
  - Geneseq: AAP60074
  - 1991-10-29
  - SUNTORY
  - EP205038-A; New tumour...

- **HUGO**
  - TNFSF9
  - TNFSF10
  - TNFSF11

- **homologs**
  - NP_000585.2 NP_036807.1 | RAT
  - NP_000585.2 NP_038721.1 | MOUSE
  - NP_000585.2 XP_858423.1 | CANFA

- **GO**
  - Function
  - transcription
  - initiation
  - elongation

- **aliases**
  - TNFA HUMAN
  - Q1XHZ6
  - IPI00001671.1
  - INCY: 1109711.Fl1p
  - CCDS4702.1
  - gi: 25952111

- **sequences**
  - >Unison: 98
  - MSTESMIRDVE...FGIIAL
  - >Unison: 23782
  - VRSSRTPSD...FGIIAL

- **SNPs**
  - P84L
  - A94T

- **protein features**
  - 1 23 SS
  - 108 143 1.8e-06 EGF
  - 162 184 TM
  - 133 138 ITIM

- **taxonmies**
  - 9606 Homo sapiens
  - 10090 Mus musculus
  - 10028 Rattus rattus

- **loci**
  - 1 23 3 6+: 31651498-31653288

- **alignments**
  - TNFA 1tnfA
  - TNFA 1tnfB
  - ... TNFA 5tswF

- **protein features**
  - MSTESMIR
  - DVEFGIIA
  - TESMIRDV
  - ITAMDAC

- **aa-to-resid**
  - MSTESMIR
  - DVEFGIIA
  - TESMIRDV
  - ITAMDAC

- **probes**
  - HGU133P
  - WHG

- **structures**
  - 1tnf
  - 1a8m
  - 2tun
  - 4tsv
  - 5tsw

- **SCOP**
  - all alpha
  - all beta
  - Ig
  - TNF-like
  - alpha+beta
Ex1: Mine for sequences w/conserved features.

- **Patents**: Geneseq:AAP60074 1991-10-29 SUNTORY EP205038-A; New tumour...
- **HUGO**: TNFSF9 TNFSF10 TNFSF11
- **Homologs**:
  - NP_000585.2 NP_036807.1 | RAT
  - NP_000585.2 NP_038721.1 | MOUSE
  - NP_000585.2 XP_858423.1 | CANFA
- **GO**: Function transcription initiation elongation
- **Aliases**: TNFA HUMAN Q1XHZ6 IPI00001671.1 INCY:1109711.FL1p CD5S4702.1 gi:25952111
- **Sequences**:
  - >Unison:98 MSTESMIRDVE...FGIIAL
  - >Unison:23782 VRSSSRTPSD...FGIIAL
- **SNPs**: P84L A94T
- **Entrez**: gene_id symbol locus
- **Taxonomy**: 9606 Homo sapiens 10090 Mus musculus 10028 Rattus rattus
- **Probes**: HGU133P WHG
- **Genomes**: Hs35 Hs36 RAT
- **Alignments**: TNFA 1tnfA TNFA 1tnfB ...
  - TNFA 5tswF
- **Protein Features**:
  - 1 | 23 | 1.8e-06 | SS
  - 108 | 143 | 1.8e-06 | EGF
  - 162 | 184 | 1.8e-06 | TM
  - 133 | 138 | 1.8e-06 | ITIM
- **Loci**: 1 233 6+:31651498-31653288
- **Probes**: 1tnf 1a8m 2tun 4tsw 5tsw
- **Structures**:
  - TNFA 1tnf 2tun 4tsw 5tsw
  - TNFA 1a8m
- **SCOP**: all alpha all beta Ig TNF-like alpha+beta
- **Taxonomy**:
  - 9606 Homo sapiens
  - 10090 Mus musculus
  - 10028 Rattus rattus
Answer more sophisticated questions.
  • Require orthologs or a specified exon structure.

Annotate hits.
  • Annotate with locus, probes, HUGO gene name, structures, PubMed refs, external links.
  • Group splice forms by locus.

Explore alternatives.
  • How do parameters influence results?
  • Try other prediction algorithms.

Stay current.
  • When new data are available, just rerun the query.

Move on.
  • The same data are available to other projects and other people.